Experimental Analysis of Public Policies

Sessions 2–3 Continuation

Matching Diagnostics, Randomised Experiments, and Power

Etienne Dagorn
University of Lille

Where We Stopped Last Week

Last week ended at the matching idea of common support:

$$0 < \Pr(D_i = 1 \mid X_i) < 1.$$

- We had already defined potential outcomes, estimands, selection bias, controls, omitted-variable bias, and matching.
- Matching only speaks credibly where treated and untreated units have comparable observed characteristics.
- The open question is how to diagnose the comparison, and what to do when comparability cannot be defended from observables alone.

Today's Arc

1. Finish matching: balance, workflow, and limits.
2. Move from argued comparability to randomised comparability.
3. Design an experiment: assignment unit, protocol, outcomes, and ethics.
4. Protect the design against implementation threats.
5. Read precision, power, and minimum detectable effects.

> **Guiding question**
> When can a comparison be read as a counterfactual rather than as another descriptive gap?

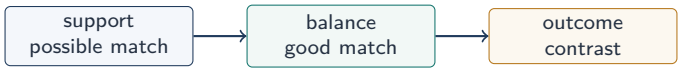
From Common Support to Balance

Common support asks

- Do comparable units exist?
- Are treated workers compared to plausible controls?
- Should part of the target population be trimmed?

Balance asks

- Did matching actually make X_i similar?
- Which covariates remain imbalanced?
- Is the final comparison credible enough to interpret?



Balance Is Checked Before Interpreting Outcomes

Before matching

- Treated workers have longer unemployment spells.
- Control workers include many low-risk cases.
- The outcome gap mixes treatment and baseline differences.

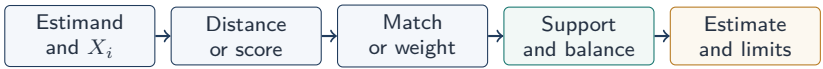
After matching

- Covariate distributions should look similar.
- Standardised differences should shrink.
- Balance on X_i is not balance on unobservables.

Order of operations

Design the match, check support and balance, then interpret the outcome contrast.

Matching Workflow



- The workflow is a design exercise before it is an outcome analysis.
- A poor balance check sends us back to the score, distance, caliper, or target population.

Regression and Matching Solve the Same Problem Differently

	Regression	Matching
Comparison	adjusted conditional mean	explicitly similar workers
Main choice	controls and functional form	distance, score, weights, caliper
Diagnostic	specification and sensitivity	balance and common support
Shared assumption	observed covariates are sufficient for comparability	
Useful combination	match or weight first, then adjust remaining differences with regression	

Observational Adjustment Has a Clear Boundary

$$(Y_i(0), Y_i(1)) \perp D_i \mid X_i$$

Diagnostics can show

- Similar observed covariates.
- Common support.
- Sensitivity to adjustment choices.

Diagnostics cannot show

- Absence of unobserved motivation.
- No hidden networks.
- No unmeasured health constraints.

Discussion question

Which unobserved factor would worry you most in the job-search workshop example?

From Argued Comparability to Created Comparability

Observational adjustment

Comparability is argued using observed X_i.

Regression and matching ask whether covariates reconstruct the missing counterfactual.

Randomisation

Comparability is created by random Z_i.

The assignment mechanism creates groups before outcomes are realised.

observed X_i $\xrightarrow{\text{turning point}}$ random Z_i

Randomisation Breaks the Link Between Assignment and Potential Outcomes

$$Z_i \perp (Y_i(0), Y_i(1))$$

What randomisation solves

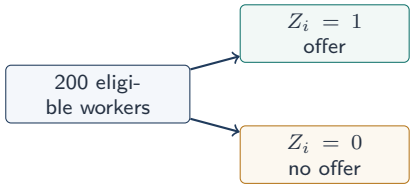
- Observed selection.
- Unobserved selection.
- Baseline comparability in expectation.

What it does not solve

- Spillovers.
- Attrition.
- Non-compliance.
- Weak power.

Workshop Lottery Creates the Comparison Group

- 200 eligible workers.
- 100 available workshop places.
- The office randomly offers access to half.
- Scarcity makes the lottery administratively meaningful.



A Difference in Means Is Causal Under Random Assignment

Assume full compliance, so $D_i = Z_i$:

$$E[Y_i(0) \mid Z_i = 1] = E[Y_i(0) \mid Z_i = 0]$$

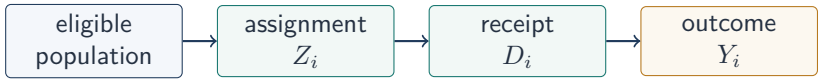
$$E[Y_i(1) \mid Z_i = 1] = E[Y_i(1) \mid Z_i = 0].$$

Therefore:

$$E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] = E[Y_i(1) - Y_i(0)].$$

The equation is justified by the assignment mechanism, not by a richer regression specification.

Assignment, Receipt, and Outcomes



$$SATE = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}$$

Full compliance: $D_i = Z_i$.

$$ITT = E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]$$

Imperfect compliance: assignment effect first.

RCT Regression Estimates Assigned-Arm Contrasts

$$Y_i = \alpha + \tau Z_i + \varepsilon_i$$

$$Y_i = \alpha + \tau Z_i + X_i'\gamma + \lambda_s + \varepsilon_i$$

- Z_i is randomised assignment.
- τ is an assigned-arm contrast.
- X_i can improve precision.
- λ_s reflects stratified assignment.
- With imperfect compliance, τ is an ITT.

Experimental Protocol Checklist

1. Eligibility rule.
2. Unit of randomisation.
3. Treatment and control conditions.
4. Assignment probabilities.
5. Primary outcome and date.
6. Strata or clusters.
7. Main estimating equation.
8. Attrition and compliance plan.

Implementation standard

Another team should be able to reproduce the assignment rule from the protocol.

Design Choices: First Tradeoffs

What do you expect?

For the job-search workshop, would you randomise workers, caseworkers, or employment offices?

Unit	Attractive when	Main risk
Worker	private treatment, little contamination	spillovers through networks
Caseworker	staff behaviour is treated	few assignment units
Office	office-wide programme	power and local shocks

Three Randomisation Designs

Design	Mechanism	When useful
Simple	randomise each eligible unit with probability p	large samples, individual delivery, low spillovers
Stratified or blocked	randomise within pre-defined groups	small samples, strong baseline predictors, local-office balance
Clustered	randomise offices, schools, villages, or regions	group delivery, contamination risk, administrative constraints

Stratification Improves Balance on Planned Variables

- Define strata before assignment.
- Randomise within each stratum.
- Common strata: office, region, unemployment duration, baseline outcome.
- Analyse consistently with the design.

Example

Randomise within local offices so treated and control workers face the same local labour market.

$$Y_i = \alpha + \tau Z_i + \lambda_s + \varepsilon_i$$

Cluster Randomisation Trades Contamination for Power

- Assign treatment to groups: offices, schools, villages, firms, or regions.
- Useful when the intervention is delivered at group level.
- Reduces contamination between treated and untreated units.
- Requires many independent clusters, not only many individuals.

Intuition

Thirty workers in one treated office provide less independent information than thirty treated offices.

Balance Tables Diagnose the Realised Assignment

Report

- sample size by arm;
- control mean;
- treatment-control difference;
- standardised difference.

Do not overread

- Balance tests do not prove randomisation.
- Small samples can have important imbalances.
- Large samples can flag tiny differences.

Diagnostic rather than proof

Validity comes from the assignment process; balance tables document the realised sample.

Outcomes and Pre-Specification Link the Estimate to the Decision

Outcome choice

- Primary outcome: employment after six months.
- Mechanism: number of applications.
- Secondary outcome: earnings after one year.

Pre-specification

- sample;
- primary outcome;
- main contrast;
- covariates and strata;
- treatment of missing outcomes.

Ethics and Feasibility Shape the Estimand

Ethical feasibility

- Scarce places can justify a lottery.
- Delayed treatment may be acceptable.
- Exclusion rules must be transparent.

Econometric implication

- Feasible assignment determines Z_i.
- Z_i determines the primary contrast.
- The contrast determines the estimand.

Randomisation Solves Selection but Not Implementation

Before implementation

$$Z_i \perp (Y_i(0), Y_i(1))$$

Treatment and control groups are comparable in expectation.

After implementation starts

The realised comparison can be threatened by spillovers, attrition, noncompliance, and measurement changes.

Implementation Threats and Responses

Threat	Problem	Design response
Spillovers	controls affected by treated workers	cluster, measure exposure, define spillover estimands
Attrition	outcome observed for a selected sample	tracking, attrition table, bounds, sensitivity
Non-compliance	$D_i \neq Z_i$	report ITT and first stage; use IV logic later
Measurement changes	outcome measured differently by arm	common instruments, timing, audit trail

Non-Compliance Is the Bridge to Session 4

Clean assignment

$$Z_i \perp (Y_i(0), Y_i(1))$$

- The offer is randomly assigned.
- The ITT remains interpretable.

Endogenous receipt

$$D_i \neq Z_i$$

- Some offered workers do not attend.
- Some controls find similar support elsewhere.

Session 4 uses Z_i as an instrument for D_i.

Bias, Precision, and Power Are Different Problems

Issue	Question	Main tool
Bias	Is the comparison causal?	design and assumptions
Precision	How uncertain is the estimate?	standard errors
Power	Would a meaningful effect likely be detected?	ex ante sample design
External validity	Where else does the result apply?	population and setting analysis

Randomisation addresses bias from selection; it does not guarantee an informative estimate.

What a Standard Error Means

$$\widehat{\tau} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$$

The estimate is one draw

- The lottery creates one treatment group and one control group.
- A different lottery would give slightly different groups.
- The estimated effect would move around across possible lotteries.

The standard error measures

- how much sampling noise remains;
- how precisely the effect is estimated;
- how much uncertainty to attach to $\widehat{\tau}$.

Mental model

Small standard error: repeated lotteries give similar estimates. Large standard error: the estimate is noisy.

Standard Errors Translate Design into Uncertainty

For a two-arm experiment:

$$SE(\widehat{\tau}) \approx \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}.$$

Term	Interpretation
σ_1^2, σ_0^2	outcome variation inside each arm
n_1, n_0	number of observations in each arm
σ^2/n	average noise left after taking a group mean

How Design Choices Move the Standard Error

Precision improves when

- n_1 and n_0 increase;
- assignment shares are balanced;
- outcome variance falls.

Precision weakens when

- assignment is clustered;
- attrition reduces sample size;
- outcomes are measured noisily.

Quick check

What happens to the standard error if we double both treatment and control sample sizes?

Minimum Detectable Effect: The Plain-Language Idea

Question before collecting data

How large would the true programme effect need to be for this design to detect it reliably?

- The MDE is a property of the design, not a result from the data.
- It combines the planned standard error, significance level, and desired power.
- It should be compared with a substantively meaningful effect.
- If the MDE is too large, the experiment may be credible but not informative.

Minimum Detectable Effect: The Calculation

$$MDE \approx \left(z_{1-\alpha/2} + z_{1-\kappa}\right) SE(\widehat{\tau})$$

Statistical choices

- α: false-positive risk.
- $1-\kappa$: desired power.
- Two-sided 5% test and 80% power: multiplier ≈ 2.8.

Design input

- $SE(\widehat{\tau})$ comes from sample size, variance, clustering, and attrition.
- Lower standard errors mechanically lower the MDE.

smaller SE $\Rightarrow$ smaller detectable effect

Numerical Reading of Power

Suppose:

- employment after six months is binary;
- $n_1 = n_0 = 200$;
- the outcome standard deviation is roughly 0.5.

$$SE(\widehat{\tau}) \approx \sqrt{\frac{0.5^2}{200} + \frac{0.5^2}{200}} \approx 0.05$$

$$MDE \approx 2.8 \times 0.05 \approx 0.14.$$

This design is built to detect effects around 14 percentage points, not small employment gains.

How to Read an MDE in a Paper

If the policy target is large

- Example: a 15 percentage-point employment gain.
- MDE $= 14$ points is close to the policy target.
- A null result is more informative.

If the policy target is small

- Example: a 4 percentage-point employment gain.
- MDE $= 14$ points is too large.
- A null result may simply mean the design was too noisy.

Reading habit

Before interpreting a null result, compare the MDE to the smallest effect that would matter for the policy decision.

Design Levers for Precision

Reduce standard errors

- more independent units;
- balanced assignment shares;
- baseline outcomes;
- strong pre-treatment predictors.

Improve the design

- stratify on outcome predictors;
- reduce measurement error;
- increase clusters when assignment is clustered;
- align analysis with randomisation.

Cluster Design Effect

$$DE = 1 + (m - 1)\rho$$

- m: average cluster size.
- ρ: intra-cluster correlation.
- When outcomes are highly correlated within clusters, effective sample size falls.
- This is why clustered RCTs need many clusters.

$$m = 25, \quad \rho = 0.05 \quad \Rightarrow \quad DE = 1 + 24 \times 0.05 = 2.2.$$

Method Map

Approach	Variation	Main assumption	Diagnostic
Naive gap	observed treatment status	treated and untreated are comparable	baseline differences
Regression	residual variation after controls	conditional independence and good controls	sensitivity
Matching	matched or weighted comparisons	conditional independence and overlap	balance, support
RCT	random assignment Z_i	assignment implemented as designed	protocol, balance

Each method is a way to justify a counterfactual comparison; none should be read as a command that automatically produces causality.

Estimand and Population Can Change Across Methods

Observational adjustment

- Regression: population depends on model support.
- Matching: trimming can focus on common support.
- ATT is often more credible than ATE when treated units define the match.

Randomised experiment

- SATE for the experimental sample.
- ITT when assignment differs from receipt.
- External validity requires a separate population argument.

Design Checklist for Reading an RCT

1. What is randomised?
2. At what level?
3. What is the estimand?
4. Is $D_i = Z_i$, or is receipt different from assignment?
5. What outcomes are measured and when?
6. What threatens implementation?
7. Is the study powered for a meaningful effect?

Transition to Session 4

Next question

What if people do not comply with their assigned treatment?

- A clean lottery first identifies the effect of assignment or offer.
- Treatment received D_i may differ from assignment Z_i.
- Session 4 separates ITT, take-up, and the local treatment effect for compliers.

References and Further Reading

- Rosenbaum and Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects".
- Imbens and Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences*.
- Duflo, Glennerster, and Kremer (2007), "Using Randomization in Development Economics Research: A Toolkit".
- Gerber and Green (2012), *Field Experiments: Design, Analysis, and Interpretation*.
- Glennerster and Takavarasha (2013), *Running Randomized Evaluations: A Practical Guide*.